

High content screening – the next challenge: effective data mining and exploration

The use of high content screening within HTS is growing and with many past hurdles now overcome, the need for effective tools for data analysis is becoming paramount.

Over the past several years, high content screening has made tremendous advances through improvements in key elements: reliable and more affordable high content readers, improved cell and liquid handling, automation of data acquisition and improved image analysis. High content screening thus has passed beyond exploratory basic research to systematic screening experiments that are incorporated into routine processes that permeate a wide range of departments at drug development organisations and academic institutions. Part of this growth is due to the incremental advancements in the field that have broadened the applications of the technology, while increasing the throughput and decreasing the overall costs. At a more basic level though, high content systems are advancing due to their unique ability to quantify the heterogeneous responses of diverse biological systems. They allow the construction of new *in vitro* assays which better reflect cellular biology and physiological context, the quantification of their response to experimental variables at unprecedented level of detail, and the automation of such measurements.

With the increasing breadth of high content screening and analysis, the field has become progressively more difficult to define. For the purposes of this article, the definition will be limited to the use of automated microscopy and image processing to measure morphology, localisation, movement, structures and organisation within cells, as determined in plate-based experiments. Although many of the tools described herein might

also be appropriate for related fields of screening such as flow cytometry and automated analysis of images from tissue sections, these are beyond the scope of this discussion.

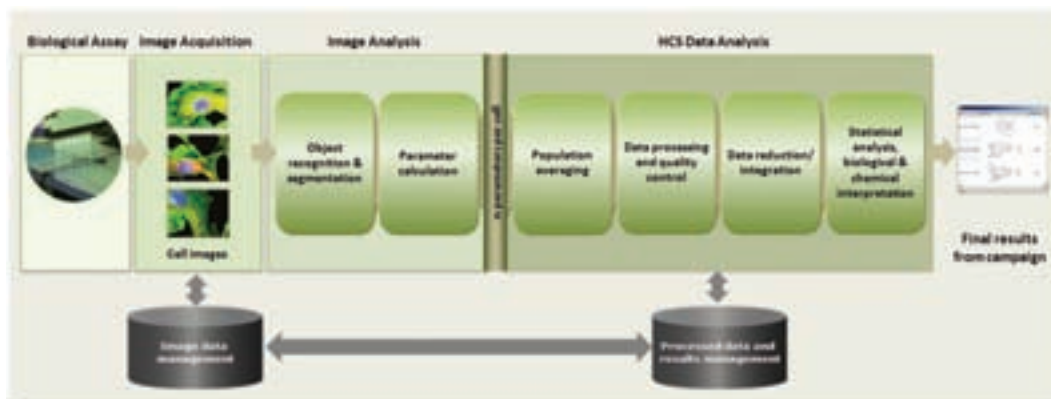
While the origins of high content screening lie in microscopy and cytology, increasing standardisation, automation and simplicity of operation drive the field in a direction where high throughput screening (HTS) found itself about 10 years ago. As such, high content screens are starting to make use of the automation and workflows of HTS programmes. While it is true that high content screening (HCS) reaches beyond HTS departments at many institutions, it is important to remember that many, but not all, of the lessons learned in HTS will be applicable to HCS. This is true from the basic issues of tools, workflows and scalability, to the more institutional issues of personnel, training and continuity of knowledge. As such, HCS now faces some of the same hurdles for throughput, storage and analysis that HTS labs faced not too many years ago when they moved from thousands to hundreds of thousands of compounds. Previous to that point, assay quality, instrument technology and its adoption had been the main challenges. Since then, however, issues of data management, scalability, results quality, reporting and decision support have moved into focus. High content screening stands at a similar transition: while assays, instrumentation and image analysis have matured, questions of scalability, data quality, data optimisation and data interpretation are today's predominant challenges.

**By Dr Kurt Zingler
and Dr Stephan
Heyse**

Figure 1

High content screening begins with setting up the biology of the assay, screening a library against it, and measuring effects via automated microscopy ('image acquisition'). Images are stored for future reference or re-analysis in an image management system. They are analysed to localise cells ('object recognition') and quantify their phenotypes in the form of numerical readouts ('parameter calculation'). These parameters then enter the data analysis work flow.

Data analysis typically encompasses the four steps of: (i) summarising the state of a cell population measured per well ('population averaging'), (ii) primary processing and quality assurance of the resulting well-based data sets ('data processing and QC'), (iii) combination of individual parameters, replicate measurements and dose-response series to a concentrated set of meaningful results quantifying effects in the screen ('data reduction'), and (iv) mining and interpretation of data across the complete HCS data set, in the context of additional chemical or biological information to answer the scientific questions the screen was designed for and obtaining the final results ('statistical analysis, biological & chemical interpretation')



This discussion will thus focus on current challenges in data management and analysis in high content screens, positioning it within the overall work flow, describing its different elements, and exemplifying data analysis in three prominent application areas of high content screening, namely target discovery, hit identification and lead development.

HCS requirements

Regardless of the particular application, High Content Screening requires the following tools for effective usage:

- Biological assays
- Automated cell and liquid handling machinery
- Screening libraries (compounds, siRNAs, or biologicals)
- Image acquisition
- Image processing
- Image storage and management
- Data processing and analysis

The first step of any high content screen is the development of the assay to measure a given biological state, or change in that state. As defined above, these high content assays generally measure the morphology, localisation, movement, structures and organisation within cells. This requires the biological system itself (the cells, expression constructs and so on) as well as the tools required for visualising responses and scoring the assay (generally a series of fluorescent probes for specific proteins, structures or cell constituents). To effectively quantify a response an assay needs robust controls to measure both the lack of a response (negative control) and a prototypical response (positive control). The assay as a whole also needs to be robust, reproducible and amenable to miniaturisation. Beyond in-house assays, there are also a wide variety of commercially available assays.

With a robust assay in hand, the next requirement is for sophisticated image acquisition hardware that is available from a number of vendors (ThermoFisher, PerkinElmer, General Electrics, BD Biosciences and Molecular Devices, to name a few). The basic task of each instrument is to automatically capture and store fluorescent images from wells of microtiter plates. The instruments vary widely in price, flexibility and throughput, but they fall into three main categories based on the type of imaging optics they use: laser scanners, bright field and confocal microscopes. Confocal instruments generally have better signal-to-background contrast, higher resolution and work better with thicker specimens, while bright field instruments are simpler, cheaper and can well be used for thin specimens (eg, a monolayer of cells). Laser scanners are very sensitive and allow very precise quantification of signals and object positions. Given the variability in the use of HCS, large research groups generally have devices from several vendors that each addresses a specific range of applications. This diversity is an important point to consider for both image storage and data processing, as an optimal system will need to consolidate data across these platforms.

In general, most of the imaging instrument providers, as well as several commercial and open-source suppliers, provide software packages for further processing the microscope images obtained. Such image analysis software generates numeric output from the images for a host of pre-defined readouts, ranging from cell number to localisation parameters for labelled cell constituents. Unlike flow cytometry, which measures just the overall fluorescence of each cell (multiple channels), the strength of HCS image analysis is its ability to recognise and quantify cell morphology, localisation, movement, structures and organisation within cells. The range of measurements is too

varied to be discussed in detail here, but assays measuring the sub-cellular localisation (or movement) of tagged molecules, neurite growth and apoptosis are common. Outside of the primary measurement, a host of other measurements are also taken on a per cell basis such as cell number, shape and size. The result is a large number of measurements on a per cell basis, which generally are integrated to per-well population averages and measures of variability.

While the acquisition hardware generally provides basic image storage capabilities, high throughput labs can generate image data at a rate that is measured in terabytes per day, and most groups doing extensive HCS work require a more sophisticated image storage and management system. These are available from a few of the same commercial vendors and at this stage they are still frequently being developed as custom in-house solutions. The basic role of these systems is to store the many terabytes of image data being captured from these instruments and to make these images easily available. Given these storage needs, solutions frequently tier the storage for faster and slower access depending on the ongoing needs for analysis. This tiered approach enables cheaper storage to be used for older, less accessed data. While not an absolute requirement for a single experiment in isolation, effective image management is essential for any dedicated HCS facility.

While most of the steps above can be completed via a host of commercially-available systems, there are only a few commercial vendors providing HCS-specific systems for the data analysis following image processing. The lack of more providers in this area may be due in part to the diversity in the field which uses a host of assays, a host of different devices and that is still in the process of developing standardised data mining techniques. This diversity within the HCS field, as well the divergence from classical HTS assays has made it more difficult for conventional HTS analysis providers to easily meet the demands in this area. This may also be simply a matter of timing, as robust and widely applicable HCS assays, instrumentation and image analysis software have only become available in the last few years. Only now the need for HCS-specific data analysis is emerging on a broad front. This topic of effective data mining and exploration is treated in the following.

Data analysis

To better understand the analysis of HCS data following image processing, it is helpful to first go through the steps in the analysis process and

then cover each in more detail. The key steps are as follows:

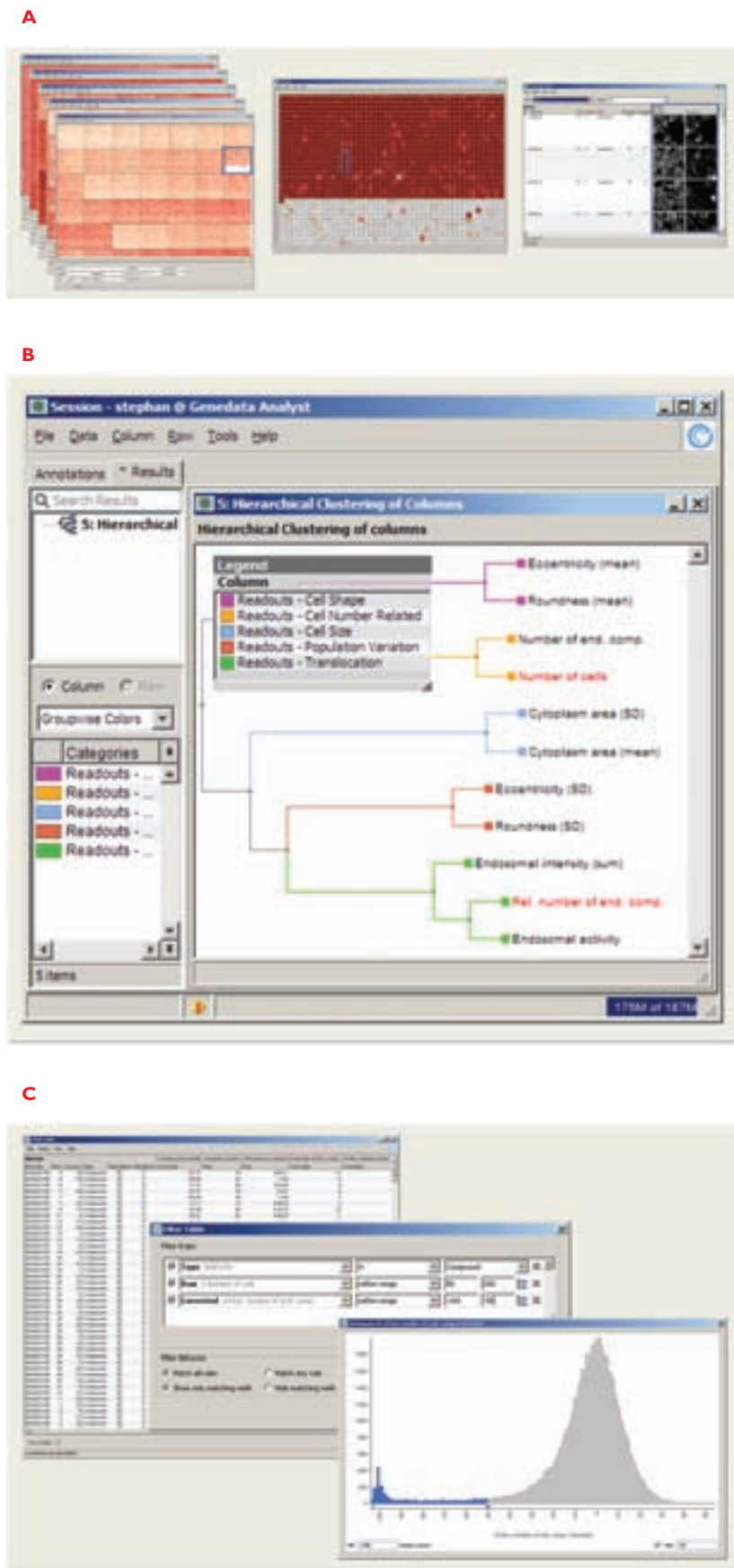
- Data loading
- Standardisation
- Quality control and assessment
- Data condensing
- Picking hits
- Mining higher-order information

While data loading is in theory a simplistic step, in practice three factors make this more problematic. First, a wide range of measurements are taken for each sample and each of these must be loaded and processed properly. Most historical HTS systems were designed to deal with a single measurement, and are overwhelmed by the new complexity in the form of HCS data. Second, given the diversity both of hardware providers and providers of image analysis solutions, there are a variety of data formats, with divergent kinds of measurements and a lack of standardisation in their nomenclature, annotation and meaning. This makes it challenging to integrate these heterogeneous outputs from individual providers into a consistent HCS data stream for systematic analysis and interpretation within an organisation. Third, while images are quantified on a per cell basis, the data are generally available as summary statistics per well (mean, median, standard deviation, etc). Some analyses are best done looking at the whole population of cells, rather than the provided summaries, and thus a method of capturing and processing cell-level data for population analysis is important.

Following capture, the system needs to be able to standardise responses to plate controls to effectively quantify the strength of a given response within each test plate and across an entire assay. The data analysis system thus needs to compare each measured well in a plate to a set of standards on that same plate. Having both positive and negative controls for this is crucial. In the simple case, these are basic numeric comparisons to obtain a percentage range for activity. In the more sophisticated case of cell population analysis, more refined algorithms, such as the Kolmogorov-Smirnov (KS statistic), are used to compare sample wells to controls.

With standardisation in place, the next step is to assess the overall quality of the assay. Effective tools will make use of Z' scores and look for plate-to-plate and within-plate trends that are unexpected. Where there are problems the data must be corrected or the assay rerun, and quality scores should be captured on the way.

If the assay in question is looking for active



compounds, the next step in the process is to pick hits using the available data. In classical HTS picking hits was done using a single measurement per well to designate the most active compounds. While this can also be done in HCS, the most effective strategies for picking active compounds involve making use of many of the available readouts in parallel, for higher selectivity of hits and better hit confirmation rates.

If applicable, data needs to be condensed across simple replicates to look for average signal and variation, and for an inhibition/activation constant in the case of dose-response curves. These steps also need careful assessment for data quality.

Finally, HCS enables the researcher to mine for higher order information from a given assay, beyond picking hits of quantifying an inhibition constant. Depending on the intent of a given assay, this information might be an added value to the work or the primary goal. An example of this kind of data mining is to look for relationships between the various HCS read-outs, such as cell number or neurite length, to see if there is a correlation in the activity that might be important in subsequent analysis. Using factor analysis, this information can be used to better understand the basic biology of a given system and the relationship between various biological states.

Beyond the requirements for the individual steps above, it is important that any data analysis system

Figure 2

A Data from a high content screen are a complex multitude of read-outs for each individual measurement, adding a new dimension as compared to conventional screening (left). They offer additional options for quality control, eg monitoring the number of viable cells for each individual well. In plate context, systematic quality problems become apparent (middle, reduced cell number in the lower section of the plate). Interactive reference to the original microscope images is very helpful for pinpointing quality issues, but also interpretation of biological effects reflected in the read-outs (right)

B The complex set of HCS read-outs can be automatically categorised by hierarchical clustering on the set of screening data, leading to distinct groups of correlated readouts. Taking the assay biology into account, these criteria are validated and annotated, and form the basis for subsequent selection of read-outs for hit filtering

C Providing more information than conventional screens about the effects of compounds or siRNA in the assay, HCS allows a more targeted selection of hits. In a simple case, two read-outs (here: cell number and a translocation measure) are combined to a filter chain with individual pass criteria, resulting in compounds which are active on the target pathway, have been measured at the right cell density, and do not kill the cells

for HCS be scalable, flexible, appropriate and integrated. It needs to be scalable, because of the additional layers of information provided by HCS that multiply the data volume being analysed. An HCS assay with 10 measured read-outs and one million compounds will have 10 times more information to process than an HTS experiment with the same number of compounds. The system needs to be flexible to meet the varied demands for data processing and analysis, and to enable a researcher to fully explore the value in the added content. It needs to be appropriate to encompass the breadth of the possible analysis steps above. It also needs to be integrated to be able to bring in data from a range of the vendors listed above, to bring in related information that helps in the analysis of the experiment (pathways, microarrays, proteomics, etc), and to smoothly output intermediate data and final results to corporate data warehouses. In short, the system needs to address a broad range of functions to fully recognise the value of HCS for the individual researcher and for the organisation. Examples of three application areas for analysis of HCS are described in more detail below. This is not an exhaustive list of potential applications.

Target discovery

There has been a tremendous amount of recent work using HCS in conjunction with RNA interference (RNAi) to discover and validate gene targets for subsequent lead identification. While use of RNAi was initially done using a single gene, advances in this area have enabled this technology to screen thousands of genes in a single assay. These can be used as screens for key gene families (eg druggable targets), for genes identified in relevant transcriptional profiling experiments or in whole genome screens depending on the need. Using this technique, researchers can look for specific phenotypes that are created by the decreased expression of a given gene after introducing its complementary miRNA or siRNA construct. This phenotype and gene target in turn can form the basis of a classic compound screen for the phenotype in question.

As in standard compound screening, researchers are looking to identify the most potent effectors of a given phenotype, and they need analysis tools for standardisation, rigorous quality assessment and selection of hits to facilitate this process. Although smaller in scale than million-compound screens, the accuracy and the elimination of false positives and negatives remain crucial as artifacts can drastically skew the results and interpretation.

Unlike traditional screens, the relationship

between the genes being assayed cannot be compared using standard SAR tools. These relationships instead need to be assayed using biological activity, pathways and other gene relationships. For biological activity comparisons, the breadth of measured read-outs enables the compounds to be compared using hierarchical clustering, correlation and related data mining algorithms. Additional information, such as microarray results, can also be helpful in further characterising the response and the relationship between the effector genes, and an effective analysis system should facilitate the comparison with these additional data types.

Finally, the researcher in these assays is striving to obtain the most complete biological picture for a given result to better understand subsequent chemical leads. To create this picture, an effective analysis should also enable the types of measurements used to be compared and analysed. The intent of this is to break down all of the responses measured into a limited set of biological activities. One way to do this is via a process called factor analysis. The principle underlying factor analysis for a multivariate data set is that a group of variables that are highly correlated with each other, but not with other variables, are likely to measure a common biological phenotype or trait. As a simple example for this, a screen for neurite outgrowth will likely take measurements for number of neurites, total length of neurites, average length of neurites and number of neurite branch points. Although these are all separate measurements, they are likely all related to the growth of neurites and thus supposed to all measure the same effect. However, such a screen might identify some siRNA species that cause a change in total neurite length, but do not affect the number or branch points. These species thus indicate two separate traits or factors for the cellular regulation of neurite growth. An analysis tool to identify these kinds of relationships is thus a big advantage for pinning down potential mechanisms from RNAi screens.

Hit identification

Although hit identification is the area most similar to typical high throughput screens in intent, it is a recent newcomer to the list of typical uses for HCS, as cost and throughput have been limiting factors for large HCS screens until recently. While some of the same techniques that are used for HCS target discovery are relevant here, the primary goal is to get the best hits possible: obtain best confirmation rates, reduce false positives and rescue false negatives. Techniques and tools for getting to these best hits are paramount.

The data volumes to be dealt with here can easily exceed those of classical HTS by orders of magnitude. When data are analysed on the well-level, the size of the compound deck is multiplied by the number of HCS read-outs, which typically ranges from 10 to 100. For full cell population analysis, one or two more orders of magnitude come into play, boosting data volume beyond a billion data points per screen. Thus, scalability is crucial for any data analysis system in this area.

Given the size of the screen, assay standardisation and quality control are even more important here, as the likelihood of getting false results, just by assay statistics, is very high. One way to take advantage of the larger data size here is to use the complete data set to look for irregular patterns that are likely due to artifact. Edge and incubator effects are common issues also in HCS and a mechanism for identifying and correcting these patterns might rescue otherwise dirty results. Cell number, cell viability and staining efficiency are standard quality scores for HCS assays, as are population variability measures. From these, experimental problems such as incomplete mixing, a blocked pipettor needle, or cell starvation are easily identified, which typically are problems too strong to 'correct' and thus require invalidation of corresponding results and rerunning of the particular plates. Access to quality control tools is thus especially important here.

Another way to improve upon the quality of hits obtained is to make more complete use of the information available for selecting hits than in classical HTS. There are two basic ways in which this analysis can happen.

The first way is to use information from all of the cells measured in an individual well instead of just taking the standard average values. The KS statistic is an example of this kind of more complete population analysis as it measures the maximum difference between a test and a control population, rather than the difference of averages. It has been shown to dramatically improve hit confirmation rates in test screens.

The second way is to use all suitable read-outs combined instead of a single read-out. In the neurite outgrowth example above, the number of neurites, total length of neurites, average length of neurites and numbers of neurite branch points can each be used as a measurement of neurite growth. Using classical HTS analysis tools, hits would be selected based on a single one of these read-outs. A more advanced tool would make more complete use of these HCS data, by defining criteria on several read-outs, by combining these to a desired

'profile' of a hit, or by scoring or automated classification of compounds into hits and non-hits, based on all suitable HCS read-outs. In the first case, a researcher would simply select a cutoff value for each of these measurements and apply the criteria in sequence, being more specific in hit selection. In the second case, a researcher would depict an idealised 'profile' of a hit, eg having a strong signal in each of the read-outs. This would abrogate an absolute cutoff and instead rank all compounds in relation to the idealised profile. In the third case, a researcher would make use of factor analysis and weigh the significance of each of the individual read-outs for the phenotype of interest, neurite growth, and then score the compounds using that weighting. Finally, a researcher could assume no *a priori* knowledge of the HCS read-outs and use the assay controls as a training set for a supervised learning algorithm (such as a Support Vector Machine), and then let the control data determine the best hits using weighing across all of the measured read-outs. Each of these techniques can drastically improve upon the overall efficiency of hit selection. Their use depends upon the specific screening scenario. A tool that provides most or all of these analysis options is an optimal one for hit list generation from HCS.

Lead development

Key areas of lead development that make use of HCS are compound profiling, *in vitro* toxicology and lead optimisation. For the purpose of this discussion these three applications will be handled together. Although scalability and efficient hit selection are key traits for an HCS analysis tool for hit identification, other characteristics are more critical for high content analysis in lead development. Specifically, the ability to make use of the additional HCS measurements and the comparison with related and/or historical data are the most important for lead optimisation.

While it is still key to monitor and improve upon the primary data measurements, eg neurite length in our example above, gaining a better understanding of all of the effects of a given compound or lead series becomes crucial at this stage. One aspect of this analysis is the ability to compare the current analysis with historical results, so a tool that enables this comparison is essential here. Another key aspect is the ability to easily analyse all of the available HCS measurements to look specifically for off-target effects. For *in vitro* toxicology studies the researcher might look at read outs for cell number, cell shape and apoptosis. For specificity analysis, say in neurite growth, the

researcher may look for off-target effects such as an overall increase in cell size or changes in cell cycle that might either bias the primary readout or have potentially problematic side-effects later in development. To validate such off-target effects and fully leverage the information from an HCS experiment, the analytical combination of HCS results with related gene expression, protein expression, metabolomics and clinical data is essential. Here, biostatistics packages which are able to treat these types of data appropriately and in combination are especially helpful.

In short, where scalability and quality control features are the primary requirements for hit identification using an HCS analysis system, the overall flexibility and ability to integrate varied data types are the principal requirements for an analysis system using high content data for lead development.

Conclusion

With the increasing sophistication and throughput of high content systems from the major vendors, the need for an effective HCS data analysis solution is becoming increasingly important. While there are a number of generic data analysis systems and HTS-specific systems available in the market today, the options become extremely limited as soon as HCS labs carefully consider both the breadth and the scale of their needs in the particular range of HCS usage scenarios that they face. With many current systems already taxed by perhaps a million data points per assay, they desire a system that can scale to smoothly process tens to hundreds of millions of data points while leveraging the complex information contained in HCS data. While many current HTS data management systems were designed in a 'one readout per well' framework, HCS researchers now require an analytical framework that can capture, process and utilise this rich content, and make it accessible and understandable to others in the organisation. Generic tools can accomplish some of these tasks, but they frequently lack the appropriate algorithms and visualisers necessary for the specific applications described above, or require highly specialised training and expertise to make use of them appropriately. Effective data mining and exploration in HCS is the next challenge, currently limiting the large-scale application of HCS. A scalable, flexible software solution designed for all of these application areas now holds the key for realising the full potential for HCS across the enterprise.

DDW

Dr Kurt Zingler is the Head of US Business for Genedata. He completed his PhD and post-doctor-

al work at the University of California at San Francisco. Dr Zingler worked first as a technical specialist and later as the business lead at a number of bioinformatics companies before joining Genedata in 2004.

Dr Stephan Heyse, Head of Genedata's Screener Business Unit, was awarded his doctorate in biophysics from the Swiss Federal Institute of Technology for developing and applying optical waveguides for label-free measurement of protein-protein interactions in cell signalling. Joining Genedata in 2000, Dr Heyse initiated the Screener system for enterprise processing, management and analysis of high throughput and high content screening data. He has guided the evolution of the product in close collaboration with key partners in the pharmaceutical industry. Prior to joining Genedata, he headed a pharmaceutical high-throughput screening laboratory at Bayer AG in Wuppertal (Germany), optimising screening processes and data analysis.